# Um *corpus* anotado de construções com verbo-suporte em Português

*Amanda Pontes Rassi*[a]
*Jorge Baptista*[b]
*Oto Araújo Vale*[c]

**Resumo**

*As construções com verbo-suporte (CVS) são um tipo de construção nominal, em que o predicador central é o nome, chamado de nome predicativo (Npred), e este é auxiliado por um verbo, chamado verbo-suporte (Vsup). A abordagem utilizada para a descrição e formalização das CVS, neste artigo, é o Léxico-Gramática. Tendo em vista as diferenças sintáticas e semânticas das CVS em relação a outros tipos de construções, o objetivo deste artigo é apresentar a metodologia e os resultados da construção de um corpus anotado com construções de Vsup e de Npred. Foi construída uma lista com 4.668 CVS, considerando-se 45 variantes de Vsup e cerca de 3.200 Npred diferentes. A partir dessa lista de CVS, foram extraídas 121.198 frases do corpus PLN.Br Full, das quais foi anotada e revista manualmente uma amostra com 2.646 frases, que constituem o corpus de referência para o tratamento de CVS em Português. Esse corpus de referência poderá ser utilizado como golden standard para avaliar tarefas automáticas de identificação, extração ou classificação de CVS ou ainda para outras aplicações de Processamento Automático de Língua Natural (PLN).*

***Palavras-chave:*** *verbo-suporte, nome predicativo, Léxico-Gramática, anotação de corpus.*

*Recebido em 27/01/2015*
*Aprovado em 13/04/2015*

[a]Universidade Federal de São Carlos – UFSCar, Universidade do Algarve – Ualg. amandarassi85@hotmail.com
[b]Universidade do Algarve – Ualg. jbaptis@ualg.pt
[c]Universidade Federal de São Carlos – UFSCar, Université Catholique de Louvain – UCL. otovale@ufscar.br

## 1. Introdução

As construções com verbo-suporte (*CVS*) são formadas, a rigor, por um verbo que funciona como verbo-suporte (*Vsup*) e um nome que funciona como nome predicativo (*Npred*). Uma mesma forma verbal, como por exemplo o verbo *dar*, pode ser o predicador em construções dativas (*e.g. Rui **deu** um livro à Ana*) ou pode funcionar como uma espécie de verbo auxiliar em construções com verbo-suporte (*e.g. Rui **deu** um beijo na Ana*). No mesmo sentido, uma forma substantival pode funcionar ora como nome concreto, argumento de um predicado verbal (*e.g. Rui guardou a **proposta** na gaveta*) ou pode funcionar como nome predicativo, em um predicado nominal (*e.g. Rui fez uma **proposta** à Ana*).

Nas *CVS*, o elemento predicador central é o *Npred*, definido como um substantivo abstrato, eventivo, que impõe as restrições de seleção dos argumentos (sujeito e complementos essenciais) da construção. O verbo-suporte, por sua vez, é uma espécie de auxiliar nominal que serve apenas para veicular as marcas gramaticais de tempo-modo-aspecto e número-pessoa, as quais não podem ser veiculadas pelo nome, devido à sua própria morfologia. As relações sintático-semânticas que se estabelecem entre o verbo e os outros constituintes da construção, sobretudo com relação ao nome predicativo, são diferentes das que se observam em uma construção com verbo pleno e podem ser determinadas por diversas propriedades formais (sintáticas), experimentalmente reprodutíveis.

Diferentes testes podem ser usados para identificar as propriedades definitórias das *CVS* (RANCHHOD, 1990; BAPTISTA, 2005). O principal teste, que representa uma propriedade necessária e suficiente, é a estreita relação entre o *Npred* e o sujeito da construção (*e.g. Pelé deu um chute na bola*, interditando a construção **Pelé deu um chute do Neymar na bola*[1]). Essa relação é da mesma natureza semântica que a relação existente entre o verbo e seu sujeito, em um predicado verbal (*Pelé chutou a bola*). Por essa propriedade, alguns verbos, tais como *conceder auxílio*, *aplicar golpe*, *meter chute*, *apresentar desculpas* e outros, podem ser considerados *Vsup*. Essas variantes estilísticas apresentam propriedades idênticas às dos *Vsup* elementares (*dar*, *estar* Prep, *fazer*, *haver*, *ser* (*de*), *ter*) e são determinadas pelo nome predica-

[1] A frase seria aceitável com a seguinte interpretação: *Pelé deu o chute que o Neymar deveria ter dado na bola* (ou com o adjunto adverbial *em vez de/ no lugar de*). Note-se, contudo, que a *CVS* continua presente na oração relativa.

tivo, que as seleciona para seu suporte. Ao contrário dos *Vsup* elementares com uma distribuição mais lata e semanticamente mais neutros, estas variantes têm uma distribuição mais estrita e veiculam diferentes matizes aspectuais e estilísticos, constituindo um importante recurso expressivo da língua.

Além desse teste, há outros que podem ser indicativos de *CVS*, tais como: (i) a substituição da construção com *Vsup* por um verbo pleno correspondente (*e.g. dar um abraço* = *abraçar*, ou *dar um beijo* = *beijar*); (ii) as restrições sobre os determinantes (*e.g. Ana deu uma passeada no parque*, interditando a construção **Ana deu minha passeada no parque*); (iii) a descida do advérbio, que permite que um advérbio modificador de um verbo numa construção verbal "desça", sob a forma de adjetivo correspondente, para a posição de modificador do nome predicativo na construção nominal equivalente (*e.g. Rui chutou fortemente a bola* = *Rui deu um chute forte na bola*); dentre outros testes. Para uma visão mais geral sobre *CVS*, veja-se, dentre outros, Gross (1981, 1998), Giry-Schneider (1978, 1987), Meunier (1981), Vivès (1983), Ranchhod (1990) e Baptista (2005).

As *CVS* podem admitir transformações específicas, tais como a conversão, a passiva, a simetria ou ainda a possibilidade de formar construções causativas.

A conversão (GROSS, 1982, 1989) é uma operação formal (ou transformação) que estabelece uma relação não-orientada de equivalência sintática e semântica (parafrástica) entre duas frases elementares. Essa operação sintática "executa a permuta dos argumentos em torno do núcleo predicativo da frase sem alterar seu significado global, é semelhante à passiva das construções verbais" (BAPTISTA, 2005, p.184). A construção *standard* tem orientação ativa (*O Rui deu um esclarecimento à Ana*), em que o agente da ação se encontra alinhado com a função sintática de sujeito, enquanto a construção conversa possui orientação passiva (*A Ana recebeu um esclarecimento por parte do Rui*), estando na posição de sujeito o tema, o paciente ou o objeto da ação, enquanto o agente fica alinhado na posição de complemento preposicional da construção.

A transformação passiva já foi longamente descrita nas gramáticas e, por isso, não será explicada em pormenor. Nas construções nominais, a transformação passiva funciona da

mesma forma como nas construções verbais: a frase está na voz ativa quando o sujeito é o agente da ação (*O Rui deu um esclarecimento à Ana*) e está na voz passiva quando o sujeito da frase é o tema, paciente ou objeto – neste caso, o nome predicativo (*O esclarecimento foi dado por Rui à Ana*).

Já a simetria (BAPTISTA, 2005) é uma propriedade de certas construções em que se verifica que dois argumentos de um predicado desempenham, relativamente ao núcleo predicativo, o mesmo papel semântico. Por essa razão, os argumentos podem trocar de posição relativa e/ou serem coordenados, sem que isso altere o significado das frases resultantes, *e.g. O Rui fez um acordo com a Ana = A Ana fez um acordo com o Rui = O Rui e a Ana fizeram um acordo.* A simetria observa-se tanto em construções nominais, como no exemplo acima, como em construções verbais e adjetivais.

Além dessas propriedades transformacionais, refira-se também à possibilidade de construir frases causativas a partir de *CVS* elementares, envolvendo um verbo com um estatuto gramatical particular, a que Gross (1981, p.23) chamou *verbo-operador causativo* (*VopC*). Grande parte das *CVS* que admitem *ter* como *Vsup* das respectivas construções de base autorizam igualmente a formação de construções causativas com os verbos-operadores causativos *dar* e/ou *fazer*. Por exemplo, considera-se que a construção *A Ana tem um grande medo do escuro* está na base de frases causativas como *O incidente traumático deu na Ana um grande medo do escuro.* As construções causativas não são consideradas construções de base porque podem ser desdobradas em dois predicados semânticos: um que exprime um *estado* (*ter medo*) e outro que veicula uma relação semântica de *causa* (*dar*), ligando o constituinte que expressa essa causa (*incidente*) e a frase de base.

Neste trabalho, foram adotados os princípios teórico-metodológicos do modelo do Léxico-Gramática (GROSS, 1975, 1981), segundo os quais a unidade mínima de análise linguística é a frase simples, que contém um predicado semântico de base e seus argumentos (sujeito e complementos essenciais). Para caracterizar uma construção ou fenômeno linguístico, é preciso descrever sistematicamente suas propriedades estruturais, distribucionais e transformacionais, pois verifica-se que, apesar de certas regularidades, cada unidade lexical de

uma língua possui uma gramática própria. Tal questão leva à adoção de uma abordagem descritivista e taxonômica, em que se procura sistematicamente contrastar as propriedades dos itens lexicais e organizar os dados em conjuntos empírica e teoricamente consistentes.

## 2. Recolha dos candidatos a *CVS*

Os pares de *Vsup* e *Npred* candidatos a *CVS* foram extraídos de três matrizes do Léxico-Gramática (BARROS, 2014; SANTOS-TURATI, 2012; RASSI et al., 2014) e somam cerca de 5.800 *Npred*. Todos esses dados foram descritos nos moldes do Léxico-Gramática (GROSS, 1975, 1981), em matrizes binárias: nas linhas, constam as entradas lexicais – neste caso, os nomes predicativos; e, nas colunas, as propriedades sintáticas das construções. As propriedades sintáticas descritas nas três matrizes (uma para cada verbo-suporte) são: (i) propriedades estruturais (número de argumentos, classe sintática, tipo de preposição que introduz os complementos essenciais e tipo de determinante selecionado pelo *Npred*); (ii) propriedades distribucionais (tipo de sujeito e tipo de complemento: humano, não-humano, parte do corpo, nome locativo, oração completiva (finita ou infinitiva), oração factiva, *etc*.); e (iii) propriedades transformacionais (simetria, conversão, possibilidade de formação de grupo nominal, redução do *Vsup*, variantes de *Vsup, etc*.). Na intersecção de cada linha com cada coluna, marca-se "+" ou "–" para indicar respectivamente a aceitabilidade ou inaceitabilidade da propriedade para essa dada construção. Esse formalismo, além de facilitar a visualização, a comparação e o tratamento dos dados, também pode ser usado em aplicações de Processamento Automático de Língua Natural (PLN).

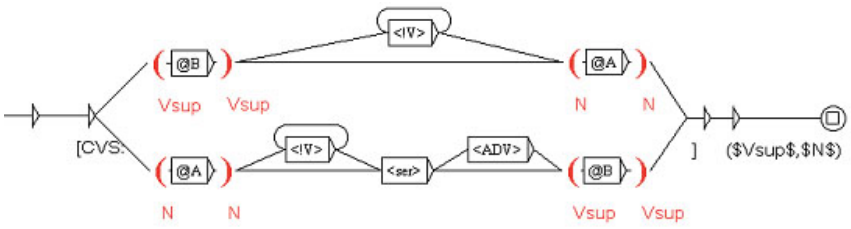
Foram recolhidas, ao todo, 4.668 *CVS* diferentes, considerando-se 45 variantes dos *Vsup* elementares *ter*, *fazer* e *dar* e cerca de 3.200 nomes predicativos diferentes. Essa lista constitui um inventário de *CVS* que pode ser usado para fins computacionais ou servir como ponto de partida para futuros trabalhos que visem a identificar ou extrair automaticamente *CVS* de *corpora*. A seção seguinte apresenta a metodologia usada para extrair esses pares candidatos a *CVS* a partir de *corpus*.

## 3. Extração dos candidatos a *CVS* a partir de *corpus*

Das 3 matrizes descritas na seção anterior, foram utilizadas as informações de verbo-suporte elementar, variantes de verbo-suporte e verbo-suporte converso, além da coluna dos nomes predicativos. A fim de identificar todas as construções em *corpus*, recorreu-se ao software Unitex[2] (v. 3.1) (PAUMIER 2003), uma plataforma *open-source* de desenvolvimento de recursos linguísticos e processamento automático de texto, baseada em tecnologia de máquinas de estados finitos. As informações sobre a co-ocorrência dos verbos com os nomes de cada construção foram, posteriormente, intersectadas com grafos de referência[3].

Assim, foram criados dois grafos de referência: (i) um a ser utilizado com os verbos-suporte *fazer* e *dar* (Fig. 1), porque leva em consideração a forma passiva das construções; e (ii) outro grafo a ser utilizado com os *Npred* que selecionam o verbo *ter*, que não considera a forma passiva porque as construções com *ter* não admitem apassivação.

**Figura 1** – Grafo de referência para identificação das *CVS* com os verbos *fazer* e *dar*




As notações e códigos usados nesse grafo estão explicadas na nota *iii* (ao fim do texto). Esse grafo possui dois caminhos. A partir da etiqueta "[CVS:", o primeiro caminho se lê: instancia-se o verbo-suporte, seguido imediatamente de um nome predicativo ou seguido de qualquer número de palavras, desde que não sejam verbos; em seguida instancia-se o nome predicativo e fecha-se a etiqueta *CVS* com "]". No segundo caminho, que prevê a inversão da ordem de verbo e nome, a fim de dar conta dos casos de construções passivas, lê-se:

[2] Disponível em <http://www-igm.univ-mlv.fr/~unitex/>

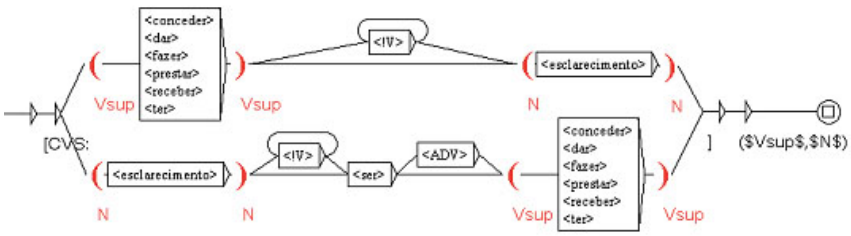
[3] Os grafos são um tipo de representação formal utilizado em abordagens estruturais para a descrição de línguas e é também um recurso disponível na ferramenta Unitex, utilizada neste trabalho. Os grafos são autômatos de estados finitos e são lidos da esquerda para a direita: a seta mais à esquerda indica o primeiro estado do grafo e o quadrado dentro de um círculo (mais à direita) indica o estado final. Entre os estados inicial e final, existem vários estados intermediários, que são representados pelas caixas (retângulos horizontais). Os caminhos entre um estado e outro são indicados por meio de setas. Nos grafos que se seguem, os estados intermediários significam: (i) "@A" e "@B" são variáveis que serão preenchidas por *Npred* e *Vsup*, respectivamente, a partir da informação constante numa matriz; (ii) "<!V>" indica a negação da categoria *verbo*, ou seja, naquela posição pode haver qualquer palavra que não seja um verbo. Na caixa <!V>, uma das setas que saem indica um loop, ou seja, a possibilidade de ocorrer uma sequência indefinidamente longa de elementos que não sejam verbos; (iii) "<ser>" corresponde a qualquer forma conjugada do verbo *ser*; e (iv) "<ADV>" corresponde a qualquer advérbio. Os símbolos "N" (nome) e "Vsup" (verbo-suporte) em vermelho identificam as variáveis usadas na saída (*output*), nomeadamente as balizas e as etiquetas "[CVS: ... ] ($Vsup$,$Npred$)" a ser visualizada no *output*, que é a saída do sistema. No caso deste grafo (Fig. 1), o *output* será, por

exemplo: *"O catálogo da exposição de Duchamp no Palazzo Grassi de Veneza, encerrada em julho último,* [CVS: *dá as primeiras pistas*] (dá,pistas)." Trata-se de um tipo particular de grafos diretos acíclicos (*directed acyclic graphs*), que se usam para processar tabelas de dados, linha a linha, e que fazem referência às colunas dessas tabelas através de variáveis. Nesses grafos, definem-se os padrões combinatórios a procurar nos textos e, posteriormente, o sistema gera, para cada linha, o correspondente transdutor que permite procurar esses padrões nos textos e, em caso de emparelhamento, modificar o texto (inserir balizas, por exemplo).

instancia-se o nome predicativo, que pode ser seguido imediatamente do verbo *ser*[4] ou pode haver entre eles um número indefinido de palavras que não sejam verbos; o verbo *ser* pode ainda ser seguido imediatamente do *Vsup* ou haver, entre eles, algum advérbio; no fim, fecha-se a etiqueta *CVS* com "]". Foi construído um grafo semelhante (considerando apenas o primeiro caminho do grafo da Fig. 1) para as construções com *ter*, que não admitem apassivação.

Ao intersectar as matrizes com os grafos de referência, o sistema Unitex produz um subgrafo de resultado para cada linha da matriz que foi instanciada por uma combinação de *Vsup* e *Npred*. Assim, o número de subgrafos corresponde ao número de linhas das matrizes. A Fig. 2 mostra um exemplo de um dos subgrafos de resultados gerado automaticamente pelo Unitex.

**Figura 2** – Subgrafo 0398 da matriz dos *Npred* com *dar*




Nesse subgrafo, representam-se as combinações determinadas pelo nome predicativo *esclarecimento*, que pode selecionar o verbo-suporte elementar *dar* ou suas variantes estilísticas e/ou aspectuais *conceder*, *fazer* e *prestar*, na construção *standard*, e selecionar os verbos-suporte conversos *receber* e *ter*, na construção conversa. No caminho inferior do grafo, apresentam-se as construções passivas correspondentes[5].

Além de todos os subgrafos de resultados, o Unitex também gera automaticamente um grafo único de resultados, que integra todos os subgrafos. Esse grafo de resultados foi aplicado ao *corpus* PLN.Br Full (BRUCKSCHEN *et al*., 2008), que é um *corpus* jornalístico do jornal *Folha de São Paulo*, com 29.014.089 *tokens*, distribuídos por 103.080 mil textos. Somando-se os re-

[4] Os verbos *estar*, *ficar* e outros auxiliares das passivas de estado não foram considerados neste momento.

[5] A possibilidade de o verbo *ter* integrar uma construção passiva (no segundo caminho do subgrafo) deve-se ao caráter automático do processo de geração do grafo. Uma vez que este é usado para reconhecimento (e não para geração), tal não constitui um problema para os objetivos deste artigo.

sultados da aplicação das 3 matrizes, o UNITEX retornou 177.287 frases. Algumas dessas frases foram extraídas em duplicado porque, muitas vezes, o mesmo nome predicativo constava em duas ou três matrizes com as mesmas variantes de *Vsup*; portanto, excluindo-se as duplicadas, restaram 131.734 frases. As frases que foram mal segmentadas pelo sistema (por problemas de pontuação, sobretudo) somam 10.536 frases (8% dos dados) e foram excluídas do *corpus* de referência. Restaram, assim, 121.198 frases a serem anotadas.

## 4. A seleção da amostra

O universo total deste objeto de estudo consiste em 121.198 frases. Desse universo, foi selecionada uma amostra estratificada a ser anotada, que é proporcional à quantidade de ocorrências de cada par (*Vsup*, *Npred*) e, ao mesmo tempo, que leva em consideração a diversidade das combinações. A amostra foi estratificada em 3 blocos: (i) a amostra global, que recuperou pelo menos um caso de todos os pares que tenham 21 ocorrências ou mais; (ii) a amostra intermédia, que corresponde aos pares (*Vsup*, *Npred*), que tenham, no mínimo 2, e, no máximo, 20 ocorrências no *corpus*; e (iii) os *hapax legomena*, que são os pares que possuem uma única ocorrência no *corpus*.

A Tabela 1 apresenta a distribuição dos pares candidatos a *CVS* em cada bloco da amostra. Ressalte-se que o número (e porcentagem) de frases aparece apenas para a amostra global porque corresponde às frases que foram, de fato, anotadas. A amostra intermédia e os *hapax legomena* serão selecionados e anotados em trabalhos futuros. A Tabela 1 indica também o número de *types*[6] em cada bloco e sua porcentagem em relação ao total, além da porcentagem do *corpus* que é abrangido por cada bloco.

[6] O termo *type* será utilizado, neste trabalho, para referir os tipos diferentes de pares de *Vsup* e *Npred*. Já o termo *token* se refere à quantidade de instâncias do mesmo objeto. O *corpus* apresenta 121.198 *tokens* (pares de *Vsup* e *Npred* iguais e/ou diferentes), mas apenas 4.668 *types* (pares de *Vsup* e *Npred* diferentes).

**Tabela 1** - Distribuição da amostra

| *Amostra* | *nº frases* | *% frases* | *nº types* | *% types* | *Cobertura do corpus* |
|---|---|---|---|---|---|
| *Amostra global* | *2.646* | *2,18%* | *1.130* | *24,2%* | *84,85%* |
| *Amostra intermédia* | --- | --- | *2.537* | *54,3%* | *14,30%* |
| *Hapax legomena* | --- | --- | *1.001* | *21,5%* | *0,85%* |
| *TOTAL* | *121.198* | *100%* | *4.668* | *100%* | *100%* |

Como se pode perceber, a amostra global abrange 2.646 frases, ou seja, 2,18% do conjunto total (121.198 frases). Dos 4.668 *types* de pares, a amostra global conta com 1.130 *types*, o que corresponde a 24,2% dos pares diferentes e 84,85% das sentenças do *corpus*. A amostra intermédia considera outros 2.537 *types*, que equivalem a 54,3% dos *types* do universo, mas apenas 14,30% das sentenças do *corpus*. O último bloco da amostra compreende os *hapax legomena*, contendo 1.001 *hapax*, que correspondem a 21,5% do total de *types* e cobrem 0,85% do *corpus*.

Conforme apontado anteriormente, apenas o primeiro subconjunto de dados, que corresponde à amostra global, foi anotado. Nesse sentido, o conjunto de frases que constituem o *golden standard* das *CVS* em Português, ou seja, o conjunto anotado que pode ser utilizado para avaliação de diferentes métodos de identificação, classificação ou processamento de *CVS*, foi estabelecido com base somente na amostra de quase 25% do total de *types* (*Vsup*, *Npred*), o que cobre 84,85% do *corpus* total.

## 5. A anotação da amostra

A anotação da amostra global (2.646 frases) foi feita manualmente por 5 anotadores falantes nativos do Português e especialistas em *CVS*, usando a ferramenta CorpusAnnotator (SUÍSSAS, 2014). Essa ferramenta foi desenvolvida em Java e precisa de dois arquivos com extensão ".txt" para funcionar: (i) um arquivo com todas as frases a serem anotadas (uma frase por linha); e (ii) um arquivo de parametrização com todas as formas de singular e plural dos nomes predicativos, a fim de assinalar em cada frase a palavra-alvo da anotação, neste caso, o *Npred*.

A anotação consistiu em assinalar, para cada frase, um código (convencional) que corresponde ao tipo de construção sintática indicada pelo par (*Vsup*, *Npred*), que aparece entre parênteses no início de cada frase. As etiquetas disponíveis são:

**CVS-*STANDARD* –** para as construções com verbo-suporte *standard*
Ex.: (*dar, tapa*) *A Ana deu um tapa no Rui*

**CVS-CONVERSA –** para as construções com verbo-suporte converso
Ex.: (*levar, tapa*) *O Rui levou um tapa da Ana*

**VOPC** – para as construções com verbo-operador causativo

Ex.: (*dar, medo*) *O vento sombrio deu medo na Ana*

Ex.: (*fazer, medo*) *O vento sombrio fez com que a Ana tivesse medo*

**OTHER** – para qualquer outro tipo de construção (com verbo pleno, expressão fixa, ou outros)

Ex.: (*fazer, academia*) *Rui fez (= construiu) uma academia* [verbo pleno]

Ex.: (*dar, tiro*) *O governo deu um tiro no próprio pé* [expressão fixa]

Ex.: (*ter, controle*) *Rui tem Max sob seu controle* [verbo-operador de ligação]

As diferenças entre as construções *standard* e as construções conversas serão abordadas na seção 7.1. As construções causativas, formadas pelo verbo-operador causativo (Gross, 1981, pp.23-38) e um nome predicativo serão explicadas na seção 7.2. As construções com o verbo-operador de ligação (Gross, 1981, pp.30-32) serão analisadas em 7.3. Já as expressões fixas serão definidas e explicadas na seção 7.4.

As anotações foram tabuladas em 5 colunas (uma para cada anotador) e, em seguida, foram comparadas por meio da ferramenta ReCal 0.1 Alpha for3+ Coders tool[7], que foi usada para calcular a concordância entre os anotadores. A Tabela 2 indica a concordância entre cada par de anotadores.

**Tabela 2** - Concordância entre pares de anotadores

| *Anotador* | *1* | *2* | *3* | *4* | *5* |
|---|---|---|---|---|---|
| *1* | --- | *83,35%* | **84,71%** | *84,28%* | *80,39%* |
| *2* | --- | --- | *81,40%* | *80,55%* | **73,31%** |
| *3* | --- | --- | --- | *84,05%* | *78,41%* |
| *4* | --- | --- | --- | --- | *77,55%* |
| *5* | --- | --- | --- | --- | --- |

Como se pode depreender da Tabela 2, as concordâncias entre anotadores variam pouco, desde a mais baixa (73,31%), entre os anotadores 2 e 5, até a mais alta (84,71%), entre os anotadores 1 e 3. A concordância média entre todos os anotadores foi de 80,8%.

[7] Disponível em <http://dfreelon.org/recal/recal3.php>

## 6. Resultados

Após a anotação, procedeu-se à análise quantitativa dos dados. A Tabela 3 apresenta a proporção relativa à quantidade de frases com concordância total ou diferentes medidas de concordâncias parciais entre os anotadores.

**Tabela 3** - Quantidade de frases com concordância total e concordâncias parciais

| *Nº anotadores* | *% anotadores* | *Nº frases* | *% frases* |
|---|---|---|---|
| *5* | *100%* | *1.584* | *59,86%* |
| *4* | *80%* | *627* | *23,69%* |
| *3* | *60%* | *318* | *12,01%* |
| *2* | *40%* | *42* | *1,62%* |
| *---* | *Excluídas* | *75* | *2,83%* |
| *TOTAL* | | *2.646* | *100%* |

Quase 60% das frases tiveram concordância entre todos os 5 anotadores. A soma das concordâncias total e parciais pela maioria (3, 4 ou 5 anotadores) corresponde a 96%. Outras 75 frases (2,83%) foram excluídas da amostra porque foram mal extraídas ou não apresentavam contexto suficiente para que os anotadores pudessem anotar.

Se, para o *golden standard*, forem consideradas apenas as 1.584 frases cuja concordância foi total entre os 5 anotadores (100% de concordância), a distribuição, por categoria, será a seguinte (Tabela 4):

**Tabela 4** - Distribuição das frases com concordância total

| *Categoria* | *Nº frases* | *% frases* |
|---|---|---|
| *CVS-STANDARD* | *1202* | *76,4* |
| *CVS-CONVERSA* | *92* | *5,8* |
| *OTHER* | *280* | *17,8* |
| *TOTAL* | *1.584* | *100%* |

A grande maioria das frases (76,4%) correspondem a *CVS standard* (de orientação ativa). Apenas 5,8% das *CVS* são construções conversas (de orientação passiva). Outras 280 frases da amostra não correspondem a construções com verbo-suporte. Vale ressaltar que não houve nenhuma frase que tenha sido anotada como *VopC* (construção com verbo-operador causativo) pela totalidade dos anotadores.

Por outro lado, se forem consideradas todas as frases cuja concordância foi de 60%, 80% ou 100%, ou seja, tiveram concordância da maioria ou totalidade dos anotadores, o *golden standard* será constituído de 2.529 frases. As frases que tiveram apenas 40% de concordância, ou seja, foram anotadas com a mesma etiqueta por apenas dois anotadores foram excluídas do *golden standard*. A Tabela 5 apresenta a distribuição das frases com concordâncias parcial (por 3 ou 4 anotadores) e total (pelos 5 anotadores), divididas por categoria.

**Tabela 5** - Distribuição das frases com concordância entre a maioria dos anotadores

| *Categoria* | *Nº frases* | *%* |
|---|---|---|
| *CVS-STANDARD* | *1.835* | *72,6* |
| *CVS-CONVERSA* | *206* | *8,1* |
| *VOPC* | *6* | *0,2* |
| *OTHER* | *482* | *19,1* |
| *TOTAL* | *2.529* | *100%* |

Como se pode notar, a grande maioria das frases (72,6%) que compõem o *golden standard* possuem *CVS* de orientação ativa (*CVS-standard*). Apenas 8,1% das frases são de orientação passiva (*CVS*-conversa). O *corpus* de referência, anotado e revisado, possui, portanto, 2.041 sentenças, que correspondem à soma do número de frases com *CVS standard* e conversa. As 6 construções com verbo-operador causativo (*VopC*), bem como as 482 frases anotadas como "OTHER" não foram integradas ao *golden standard*, já que não correspondem a construções com verbo-suporte. Contudo, esse conjunto está disponível, na medida em que constituem instâncias negativas (contraexemplos), que podem ser utilizadas para treinar e avaliar diferentes métodos de identificação, classificação e processamento de *CVS*.

## 7. Discussão de casos particulares

A maioria dos pares candidatos a *CVS* (59,86%) conforme apresentado na Tabela 3, não geram discordância entre os anotadores quanto ao seu estatuto. Há, no entanto, alguns pares cuja discordância é considerável. Nesta seção, serão apresentados casos em que houve discordância entre os anotadores e os casos em que a discordância não foi alta, porém foi sistemática. De maneira geral, esses casos particulares não se referem a um *Npred* específico, mas a grupos de *Npred* que possuem comportamentos sintático e semântico semelhantes.

### 7.1 Construções ***standard*** e construções conversas

Essa distinção está ligada à possibilidade de as *CVS* apresentarem uma construção *standard* (de orientação ativa) e uma construção conversa (de orientação passiva), ligadas pela operação de Conversão (G. GROSS, 1982, 1989; BAPTISTA, 1997; RASSI *et al*., 2015).

Os nomes predicativos relacionados à área médica podem ser interpretados sob duas óticas diferentes: do ponto de vista do *agente* (por exemplo, o médico) ou do ponto de vista do *paciente*. Trata-se de pares como (*fazer acompanhamento*), (*fazer/realizar cirurgia*), (*fazer consulta*), (*fazer diagnóstico*), (*fazer/realizar exame*), (*fazer/realizar operação*), (*fazer terapia*), (*fazer/realizar tratamento*), e outros. Os seguintes exemplos ilustram o fenômeno:

*(1)* ***(fazer, exame)*** *O médico fez um exame no paciente – STANDARD*
*(1a) = O médico examinou o paciente*

*(2)* ***(fazer, exame)*** *O paciente fez um exame no hospital público – CONVERSO*
*(2a)* ***= (submeter-se a, exame)*** *O paciente se submeteu a um exame no hospital público – CONVERSO*
*(2b) = O paciente foi examinado no hospital público (pelo médico)*

*(3)* ***(fazer, tratamento)*** *O médico faz o tratamento dos pacientes gratuitamente – STANDARD*
*(3a) = O médico trata os pacientes gratuitamente*

*(4)* ***(fazer, tratamento)*** *Os pacientes fazem o tratamento gratuitamente – CONVERSO*
*(4a)* ***= (submeter-se a, tratamento)*** *Os pacientes se submetem ao tratamento gratuito – CONVERSO*
*(4b) = Os pacientes são tratados gratuitamente (pelo médico)*

Barros (2014, pp. 99-100) havia apontado para a distinção entre os papéis semânticos de *agente* e *paciente* ocupando a posição sujeito, o que poderia auxiliar na distinção desses casos. Identificando-se os papéis semânticos dos argumentos, seria possível identificar também as diferentes orientações de sentido (ativo ou passivo) expressas pela construção.

O fato de um mesmo par (*e.g. fazer exame* ou *fazer tratamento*) poder ser classificado ora como *CVS-standard* ora como *CVS*-conversa gerou discordância entre os anotadores. A decisão adotada, nesses casos, consistiu em verificar a orientação de sentido (ativo ou passivo) da ação, por meio da aplicação de testes: transforma-se a frase nominal em uma frase verbal, mantendo o mesmo sujeito, e verifica-se se essa frase verbal está na voz ativa ou na voz passiva. Se estiver na voz ativa, a construção nominal é *standard*; se estiver na voz passiva, a construção nominal é conversa.

Uma análise semelhante à dos *Npred* da área médica pode ser feita para alguns nomes predicativos relacionados ao ensino, tais como *fazer aula, fazer curso*, *fazer/realizar prova*, *fazer/realizar teste, fazer treino* e outros.

A expressão *dar aula* forma construções tipicamente *standard*, enquanto a expressão *ter aula* forma construções tipicamente conversas. Quando esses *Npred* são associados ao *Vsup fazer*, tanto podem formar construções *standard* quanto construções conversas: se é o professor o sujeito, quem prepara a aula para os alunos, (*fazer, aula*) é *CVS-standard*; no entanto, se é o aluno o sujeito, quem assiste à aula do professor, (*fazer, aula*) é uma *CVS*-conversa; naturalmente, só conhecendo essa informação de natureza extralinguística é possível determinar com exatidão de qual construção se trata:

(5) O professor fez (= deu) uma aula sobre fungos

(6) O aluno da graduação fez (= teve) aulas de Matemática

O par *fazer empréstimo* não pertence ao mesmo campo semântico dos nomes relacionados ao estudo, mas deve ser analisado da mesma forma: *fazer empréstimo* pode ser ora *CVS-standard* ora *CVS*-conversa. Quando se refere a emprestar algo a alguém, o par *fazer empréstimo* é uma construção *standard*; quando se refere a pegar algo emprestado de alguém, o par *fazer empréstimo* é converso.

*(7)* ***(fazer, empréstimo)*** *Em 2001 Valério fez empréstimo de R$ 250 mil para o ex-tesoureiro – STANDARD*

*(7a) = Em 2001 Valério emprestou R$ 250 mil para o ex-tesoureiro*

*(8)* ***(fazer, empréstimo)*** *Em 2001 Valério fez empréstimo de R$ 250 mil no banco – CONVERSO*

*(8a) = Em 2001 Valério pegou emprestado R$ 250 mil no banco*

O primeiro caso (7), retirado do *corpus*, é tipicamente uma *CVS-standard*, enquanto o segundo caso (8), construído a partir do primeiro, é uma *CVS*-conversa. A mesma confusão se verifica com os dois sentidos veiculados pelo verbo pleno *emprestar*. Em Português do Brasil, *emprestar* admite duas regências: (i) *emprestar para*, que tem orientação ativa, e (ii) *emprestar de*, que tem orientação passiva e significa *pegar emprestado*.

Assim como no caso dos *Npred* relacionados ao estudo, o estatuto do *Npred empréstimo* é relativamente pacífico quando associado ao verbo *dar* (*standard*) e ao verbo *ter* (converso). Essa ambiguidade se verifica apenas com relação ao *Vsup fazer*.

### 7.2 Construções com verbo-operador causativo

As construções com verbo-operador causativo (*VopC*; M. GROSS, 1981) se distinguem das construções com verbo-suporte e dos verbos plenos (ou distribucionais) por várias propriedades, de que se podem destacar como principais:

1. o preenchimento lexical da posição sujeito nas construções com verbo-operador causativo sofre fracas restrições de seleção, constituindo aquilo que M. Gross (1981) designou por uma posição de nome não-restrito (*Nnr*), ou seja uma posição sintática (geralmente sujeito) que pode ser preenchida lexicalmente por nomes de tipo humano, não-humano e até orações completivas, nomeadamente factivas, e que recebe a interpretação de *causa*, *e.g.* (*O Pedro* + *A atitude do Pedro* + *O fato de o Pedro ter feito isso* + *Isso*) *deu medo à Ana*;
2. frequentemente, o verbo-operador causativo (*dar*, *fazer* e outros) "absorve" o *Vsup* da construção de base, reestruturando-a e colocando o respectivo sujeito na posição de complemento (*e.g. Isso deu/fez # A Ana tem medo = Isso deu/fez medo na Ana*); o *VopC fazer* (mas não o verbo *dar*) pode, porém, aplicar-se à frase sem que esta seja reestruturada

e permitindo, portanto, a manutenção do *Vsup* da frase de base (*e.g. Isso fez a Ana ter medo*);

3. o verbo-operador causativo pode comutar com outros verbos que, de forma mais evidente, expressam um valor causativo, tais como *causar, provocar, fazer com, etc.* (*e.g. Isso causou medo na Ana*);
4. os verbos-operadores causativos, embora tenham um largo espectro distribucional, não têm uma distribuição própria, sendo a sua seleção dependente da construção do nome predicativo da *CVS* à qual se aplicam.

Assim, alguns *Npred*, quando combinados com o verbo *ter*, formam construções de base, em que esse verbo é um *Vsup*. Por outro lado, os mesmos *Npred*, quando combinados com os verbos *dar* ou *fazer*, devem ser analisados como constituindo construções causativas com verbo-operador, como se demonstra a seguir:

(9) ***(ter, ciúmes)*** *A Ana tem ciúmes do Rui*
(9a) *[Causativo]* = ***(fazer, ciúmes)*** (*O Rui* + *Isso*) *fez ciúmes na Ana*

(10) ***(ter, alegria)*** *A Ana tem muita alegria*
(10a) *[Causativo]* = ***(dar, alegria)*** (*O Rui* + *Isso*) *deu muita alegria à Ana*

A par dessas construções causativas mais típicas, alguns *Npred*, que geralmente formam *CVS* com o verbo *dar*, podem, em situações particulares, combinar-se com o mesmo verbo em construções que deverão ser classificadas como causativas. Esses *Npred*, tais como *dica*, *explicação*, *informação* e outros, podem admitir como sujeito um nome não-restrito (*Nnr*):

(11) ***(dar, explicação)*** *A Ana deu uma explicação ao Rui sobre esse fenômeno*
≠ *(11a)* ***(dar, explicação)*** *(?Ana + O aumento da temperatura + O desmatamento da Amazônia + O fato de os homens poluírem mais + Isso) deu uma explicação sobre o aquecimento global* (*à sociedade*)

Em rigor, a frase (11) é uma *CVS*, ao passo que a frase (11a) deve ser classificada como uma construção causativa, já que atende às propriedades elencadas anteriormente. O sujeito humano de (11a) é praticamente inaceitável nessa construção com interpretação causal. Contudo, dada a possibilidade de

ambiguidade entre a interpretação agentiva e a interpretação causal de um sujeito humano nesse tipo de construções, a frase será classificada como uma *CVS* com base nos seguintes critérios: (i) a expressão corresponde a uma *ação* (logo, o sujeito recebe o papel semântico de *agente*); e (ii) o sujeito, com interpretação *agentiva*, é também de natureza *volitiva* e a ação realizada é *intencional*, o que pode ser atestado por meio da inserção, na frase, de algum advérbio, como *de propósito, intencionalmente, propositalmente, voluntariamente, etc.* (11b):

> (11b) = A Ana, (intencionalmente + voluntariamente + propositalmente), deu uma explicação incorreta ao Rui sobre esse fenômeno

Além de o predicado *dar uma explicação* em (11b) referir-se a uma ação, também é possível inserir qualquer um desses advérbios logo após o sujeito. Note-se que esse teste não é aceitável em construções causativas, como, por exemplo, (11c):

> (11c) ***(dar, explicação)*** **(O aumento da temperatura + O desmatamento da Amazônia + O fato de os homens poluírem mais + Isso) deu, (intencionalmente + voluntariamente + propositalmente), uma explicação sobre o aquecimento global* (*à sociedade*)

## 7.3 Construções com verbo-operador de ligação

Muitos nomes predicativos, tais como *notícia, orientação, informação, explicação, opinião, solução, resposta, exemplo, definição, dica, pista, sugestão, argumento, parecer etc.*, associados ao *Vsup ter*, admitem duas interpretações diferentes, uma de sentido passivo, em (12), e outra de sentido ativo, em (13):

> (12) ***(ter, notícia)*** *Zé* ***teve*** *uma* ***notícia*** *ruim* <*quando Ana lhe contou sobre a morte do pai*>

> (13) ***(ter, notícia)*** *Zé* ***tem*** *uma* ***notícia*** *ruim* <*para dar à/para Ana*>

Os tempos verbais, em rigor o aspecto "pontual" do pretérito perfeito, em (12), e o aspecto "durativo" do presente (ou do imperfeito), em (13), permitem distinguir esses dois empregos. O exemplo (12) é menos controverso, sendo claramente considerado como uma *CVS*-conversa, já que tem como contraparte a construção *standard*:

> (12a) *Alguém* ***deu*** *uma* ***notícia*** *ruim ao Zé*

O estatuto dessa construção conversa não gerou dúvidas entre os anotadores. Em contrapartida, o mesmo par (*ter, notícia*), em (13), parece ter um estatuto especial, pois se assemelha a uma construção de orientação ativa (*e.g. Zé deu uma notícia ruim à Ana*), mas a ação não chega a se concretizar (aspecto imperfectivo).

O predicado de base em (13) é *dar uma notícia*, já que pode ser reconstituído na oração infinitiva introduzida por *para* (*e.g. Zé tem uma notícia ruim para dar à/para Ana*). De acordo com Cristina Santos-Turati (comunicação pessoal), o verbo *ter*, nesse sentido, serve apenas para ligar o argumento (*Zé*) ao predicado *dar uma notícia*. Esse argumento (*Zé*) não é novo, ele já existia na frase de base. Nessas condições, o verbo *ter*, em (13), tem um estatuto de um verbo-operador de ligação (*VopL*; M. GROSS, 1981; RANCHHOD, 1990), ligando um argumento (*Zé*) a uma frase de base, embora esse argumento já esteja presente nela. Ao mesmo tempo, ele "absorve" o verbo-suporte *dar* da construção de base. Há, no entanto, outras especificidades que distinguem (13) de outras construções com *VopL*, tais como a possibilidade da reconstituição do *Vsup* reduzido numa oração introduzida por *para*, mas essa análise não será feita em profundidade neste trabalho, remetendo para o estudo da autora (SANTOS-TURATI, 2015, em preparação).

*Zé* ***tem*** **#** *Zé* ***dá*** *uma* ***notícia*** *ruim para Ana*

(13) *Zé* ***tem*** *uma* ***notícia*** *ruim para* (*Ana* + *dar à/para Ana*)

O mesmo fenômeno pode ser observado em vários outros *Npred* associados ao verbo *ter*, tais como *dica*, *explicação*, *informação*, *opinião*, *orientação*, *pista*, *solução*, entre outros. Esses *Npred*, quando combinados com o verbo *ter*, geralmente dão origem a construções conversas, deriváveis de *CVS standard* com o verbo *dar*, como é o caso de (14) a (16). Pontualmente, porém, terão de ser classificados como construções com verbo-operador de ligação (*VopL*), como é o caso de (17) a (19) (exemplos selecionados a partir do *corpus* de referência):

(14) ***(ter, notícia)*** *Segundo Zeca, o Estado vizinho de Mato Grosso tem "quase uma dezena de usinas instaladas na bacia do Paraguai sem que se tenha* ***tido notícia*** *de um único acidente ambiental"*

(15) ***(ter, informação)*** *A delegada diz que é importante que os passageiros que sejam furtados ou roubados registrem a ocorrência na delegacia do aeroporto, para que a polícia* ***tenha*** *mais* ***informações*** *sobre o modo como os bandidos agem*

(16) ***(ter, solução)*** *A disputa entre juízes e a direção da liga, que aparentemente* ***teria*** *uma* ***solução*** *rápida, deve durar algumas rodadas*

(17) ***(ter, notícia)*** *O "The Wall Street Journal"* ***tem*** *boas* ***notícias*** *para todos vocês, ratos de sofá*

(18) ***(ter, informação)*** *A página* ***tem informações*** *sobre o clube, fotos e os nomes dos membros*

(19) ***(ter, solução)*** *Quem ousaria dizer que* ***tem*** *a* ***solução*** *para o caso?*

## 7.4 ***Npred*** com interpretação não-literal ***vs.*** expressões fixas

Houve alguma divergência entre os anotadores na classificação de pares de *Vsup* e *Npred* candidatos a *CVS* nos casos em que o *Npred* possui significado metafórico (*i.e. figurado, não-literal*). Alguns anotadores classificaram-nos como *CVS-standard* e outros como OTHER. Seguem-se alguns desses casos:

(20) ***(dar, passo)*** *Pior: poucos países estão de fato* ***dando passos firmes*** *para atingir a meta*

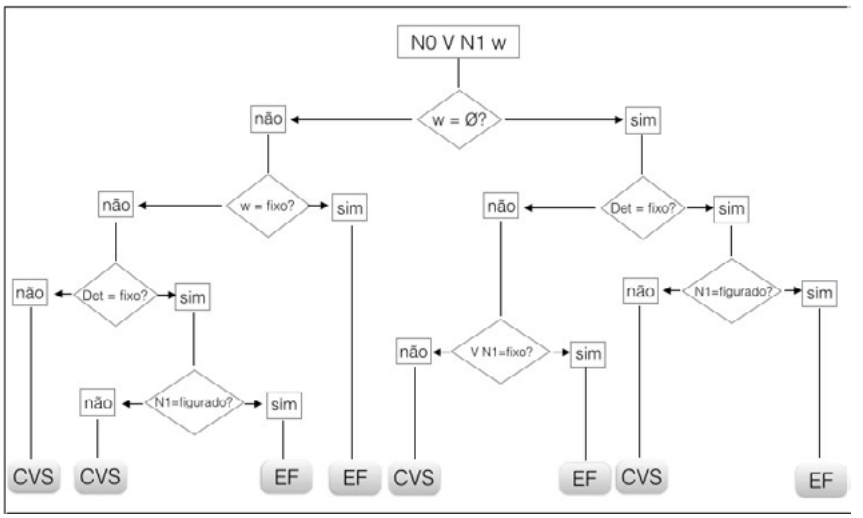
(21) ***(dar, tiro)*** *O PT está* ***dando um tiro no próprio pé*** *ao tentar abortar a CPI do caso Waldomiro*

(22) ***(dar, volta)*** *Até lá, não custa nada ter esperança de que pelo menos um grande clube carioca está* ***dando a volta por cima*** *e reconquistando seu lugar de honra na elite do futebol nacional*

(23) ***(fazer, justiça)*** *Somos revolucionários puros, a causa é nobre e a história nos* ***fará justiça***

No intuito de sistematizar a anotação, foram considerados os seguintes critérios para distinguir as Expressões Fixas (*EF*) das *CVS*: (i) a existência de complementos fixos; (ii) a fixidez do determinante; (iii) o sentido (literal ou figurado) dos complementos; e (iv) a fixidez do verbo (*V*) e do nome (*N1*), que remete para a impossibilidade de substituí-los por sinônimos. A Fig. 3 apresenta a chave dicotômica para a distinção entre *EF* e *CVS*.

**Figura 3 –** Chave dicotômica para a distinção entre *EF* e *CVS*




A Fig. 3 deve ser interpretada da seguinte forma:

1. Se a construção possui outro(s) complemento(s)[8]:
   1.1. e, se algum dos complementos é fixo, então trata-se de uma *EF*;
   1.2. mas, se o complemento não é fixo, então:
      1.2.1. se o determinante é livre, é uma *CVS*;
      1.2.2. mas, se o determinante é fixo:
         1.2.2.1. e N1 tem sentido literal, então é *CVS*.
         1.2.2.2. e N1 tem sentido figurado, então é *EF*.
2. Se a construção não possui outro(s) complemento(s):
   2.1. se o determinante é livre:
      2.1.1. e os dois constituintes (verbo e nome) são fixos, então é *EF*;
      2.1.2. mas os dois constituintes (verbo e nome) são livres, então é uma *CVS*;
   2.2. mas, se o determinante é fixo:
      2.2.1. e, se o nome tem uma interpretação literal, então é uma *CVS*.
      2.2.2. e, se o nome tem uma interpretação figurada, então é uma *EF*.

[8] A existência de outro(s) complemento(s) é indicada pela letra *w*.

Para concluir esta seção, analisem-se dois casos aparentemente semelhantes, mas que são diferentes quanto à sua classificação: (*fazer, cinema*) e (*fazer, filme*). Esses casos são ainda interessantes por salientarem as diferenças entre as variantes brasileira e europeia. No Português Europeu, além da interpretação literal de "ser ator de filmes", a expressão *fazer um filme* também tem um valor figurado quando empregue no sentido de "pôr-se a imaginar o que poderia acontecer", geralmente com uma conotação negativa (ou *disfórica*), sendo considerada, nessa situação, uma *EF*. Em Português Brasileiro, *fazer filme* permite apenas a interpretação literal, pelo que corresponde a uma *CVS*. Pode-se comparar a *EF fazer um filme*, em PE, com as expressões fixas *fazer teatro*, *fazer drama* e *fazer cena*, em PB, que, em certas situações, significam "agir com exagero".

O par (*fazer, filme*), na perspectiva tanto do cineasta quanto do ator, é considerado uma *CVS*, pois, na ausência de outros complementos, o *Npred filme* admite determinante livre. Já o par (*fazer, cinema*) também não possui outros complementos, além do verbo e do nome, então, seguindo a orientação da chave dicotômica, verifica-se se o determinante é livre ou fixo:

(24) ***(fazer, filme)*** *Eva fez (um + muitos + algum + seu + ∅) filme(s)*

(25) ***(fazer, cinema)*** *Eva fez (*um + *muitos + *algum + *seu + ∅) cinema*

O exemplo (24) admite livremente qualquer determinante, ao passo que (25) exige a ausência de determinante, ou seja, tem determinante fixo. Além da fixidez do determinante, o nome *cinema* é usado metonimicamente no lugar de "filme exibido no cinema", o que nos permite afirmar que *fazer cinema* é uma *EF*, ao passo que *fazer filme* é uma *CVS*.

Utilizando-se essa chave dicotômica, foi possível identificar 17 ocorrências de Expressões Fixas na amostra anotada, tais como os exemplos de (20) a (23).

## 8. Considerações finais

Este artigo descreveu o processo de constituição de um *corpus* anotado com informações sobre verbos-suporte e nomes predicativos, a partir da anotação manual de uma amostra das

121.198 frases extraídas de textos reais em Português do Brasil. Foram recenseadas 4.668 combinações diferentes de pares de *Vsup* e *Npred*, que podem ser usadas em tarefas de PLN, tais como análise sintática automática, desambiguação lexical, ou em diversas aplicações, como tradução automática, sumarização automática, sistemas de pergunta e resposta, dentre outras.

Além da lista de *CVS*, foi constituído também um *subcorpus* do PLN.Br Full, contendo 2.646 frases anotadas e revisadas manualmente, que servirá como *corpus* de referência para avaliar sistemas automáticos de análise sintática. No futuro, pretende-se utilizar esse *corpus* para avaliar, em especial, a performance do *parser* XIP (MOKHTAR, 2002), que é o analisador sintático automático usado na cadeia de processamento do Português STRING (MAMEDE *et al*., 2012).

## REFERÊNCIAS BIBLIOGRÁFICAS

BAPTISTA, J. Sermão, tareia e facada: uma classificação das expressões conversas dar-levar. In: *Seminários de Linguística 1*, Faro. Universidade do Algarve, Unidade de Ciências Exactas e Humanas, pp. 5-38, 1997.

_____. *Sintaxe dos Predicados Nominais com SER DE*. Lisboa: F. Calouste Gulbenkian/ Fundação para a Ciência e Tecnologia, 2005.

BARROS, C. D. *Descrição e classificação dos predicados nominais com o verbo-suporte fazer em Português do Brasil*. Tese (Doutorado em Linguística) - Faculdade de Letras, Universidade Federal de São Carlos, São Carlos-SP, 2014. Disponível em: <http://www.bdtd.ufscar.br/htdocs/tedeSimplificado/tde_busca/ arquivo.php?codArquivo=7213>. Acesso em 06/04/2015.

BRUCKSCHEN, M. *et al.* Anotação linguística em XML do *corpus* PLN-BR. *Série de relatórios do NILC*, Núcleo Interinstitucional de linguística Computacional - ICMC/USP, 2008.

GIRY-SCHNEIDER, J. *Les nominalisations en français: l'opérateur faire dans le lexique*. Genève: Librairie Droz, 1978.

_____ *Les prédicats nominaux en français*: les phrases simples à verbes support. Genève: Librairie Droz, 1987.

GROSS, G. Un cas des constructions inverses: donner et recevoir. *Lingvisticae Investigationes*, 2 (1), 1982, pp.1-44.

_____. *Les constructions converses du français*. Genève: Librairie Droz, 1989.

GROSS, M. *Méthodes en syntaxe*. Paris: Hermann, 1975.

_____. Les bases empiriques de la notion de prédicat sémantique. In: *Langages*, 15e année, n°63, 1981. pp. 7-52. doi : 10.3406/lgge.1981.1875. Disponível em: <http://www.persee.fr/web/revues/home/ prescript/article/lgge_0458-726X_1981_num_15_63_1875>. Acesso em 06/04/2015.

_____. La foction sémantique des verbes supports. *Travaux de Linguistique*, 37, 1998.

MAMEDE, N. *et al*. STRING: An hybrid statistical and rule-based Natural Language Processing chain for Portuguese. In: *Proceedings of PROPOR'12*. International Conference on Computational Processing of Portuguese (Demo session), Coimbra, Portugal, April, 2012.

MEUNIER, A. *Nominalisations d'adjectifs par verbes supports*. Tese (Thèse de Troisième cycle) - Laboratoire Automatique Documentaire et Linguistique, Université Paris 7, 1981.

PAUMIER, S. *De la reconnaissance de formes linguistiques à l'analyse syntaxique*. Tese (Thèse de Doctorat). Univ. Paris-Est, Marne-la-Vallée, 2003.

MOKHTAR, S. A.; CHANOD, J. P.; ROUX, C. Robustness beyond shalowness: incremental dependency parsing. *Natural Language Engineering*, 2002, pp.121-144.

RANCHHOD, E. M. *Sintaxe dos predicados nominais com Estar*. Lisboa: INIC - Instituto Nacional de Investigação Científica, 1990.

RASSI, A. P. *et al*. The fuzzy boundaries of operator verb and support verb constructions with dar ``give'' and ter ``have'' in Brazilian Portuguese. In: *Proceedings of Workshop on Lexical and Grammatical Resources for Language Processing*. Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University, 2014, pp. 92-101.

RASSI, A. P *et al*. Estudo contrastivo sobre as construções conversas em PB e PE. In: *Anais do I Congresso Internacional dos Estudos do Léxico e suas interfaces*, Araraquara, SP: UNESP, 2015 (no prelo).

SANTOS-TURATI, M. C. A. Descrição da estrutura argumental dos predicados nominais com o verbo-suporte *ter*. In: *Seminário do GEL* - Grupo de Estudos Linguísticos do Estado de São Paulo, 60, São Paulo, Brasil, 2012, pp. 20–21.

SUÍSSAS, G. *Verb Sense Disambiguation*. Tese (Tese de Mestrado). Instituto Superior Técnico/ INESC-ID Lisboa – Spoken Language Laboratory, Universidade de Lisboa, 2014.

VIVÈS, R. *Avoir, prendre, perdre: Constructions à verbe support et extensions aspectuelles*. Tese (Thèse de Troisième cycle), Laboratoire Automatique Documentaire et Linguistique, Université Paris 8, 1983.

**Abstract**

**An annotated corpus with support verb constructions in Portuguese**

*The support verb constructions (SVC) are a type of nominal construction, where the core predicate is the noun, called 'predicative noun' (Npred), which is assisted by a verb, called 'support verb' (Vsup). The Lexicon-Grammar theoretical and methodological framework was adopted, in this paper, for the linguistic description and formalization of SVC in Portuguese. Considering the syntactic and semantic differences between SVC and other types of constructions, the purpose of this paper is to present the methodology and results of creating a corpus annotated with Vsup and Npred. A list with 4,668 SVC was built, considering 45 variants of Vsup and around 3,200 different Npred. Based on this list, we extracted 121,198 sentences from PLN.Br full corpus, from which 2,646 sentences have been manually annotated. This sample may constitute a reference corpus for the processing of SVC and used as a golden standard for evaluating the automatic tasks of identification, extraction or classification of SVC, as well as for other Natural Language Processing (NLP) applications.*

**Keywords***: support verb, predicative noun. Lexicon Grammar, corpus annotation.*